(Translation)

PATENT OFFICE
JAPANESE GOVERNMENT

This is to certify that the annexed is a true copy of the following application as filed with this Office

Date of Application: September 30, 1999

Application Number: Japanese Patent Application No. 277918/1999

Applicant(s): Hitachi Software Engineering Co., Ltd.

August 18, 2000

Commissioner,
Patent Office: Kozo Oikawa (seal)

Certificate No. 2000-3066086

日本国特許庁
PATENT OFFICE
JAPANESE GOVERNMENT

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出願年月日
Date of Application: １９９９年　９月３０日

出願番号
Application Number: 平成１１年特許願第２７７９１８号

出願人
Applicant (s): 日立ソフトウエアエンジニアリング株式会社

２０００年　８月１８日

特許庁長官
Commissioner,
Patent Office 及川耕造

出証番号　出証特２０００－３０６６０８６

【書類名】　特許願

【整理番号】　SK11A031

【提出日】　平成11年 9月30日

【あて先】　特許庁長官　殿

【国際特許分類】　G06F 19/00

【発明の名称】　遺伝子発現パターン表示方法および装置

【請求項の数】　5

【発明者】

【住所又は居所】　神奈川県横浜市中区尾上町６丁目８１番地　日立ソフトウエアエンジニアリング株式会社内

【氏名】　野崎　康行

【発明者】

【住所又は居所】　神奈川県横浜市中区尾上町６丁目８１番地　日立ソフトウエアエンジニアリング株式会社内

【氏名】　中重　亮

【発明者】

【住所又は居所】　神奈川県横浜市中区尾上町６丁目８１番地　日立ソフトウエアエンジニアリング株式会社内

【氏名】　渡辺　恒彦

【特許出願人】

【識別番号】　000233055

【氏名又は名称】　日立ソフトウエアエンジニアリング株式会社

【代理人】

【識別番号】　100083552

【弁理士】

【氏名又は名称】　秋田　収喜

【電話番号】　03-3893-6221

【手数料の表示】

【予納台帳番号】　014579

【納付金額】　　21,000円

【提出物件の目録】

【物件名】　　明細書　1

【物件名】　　図面　1

【物件名】　　要約書　1

【プルーフの要否】　　要

【書類名】　明細書

【発明の名称】　遺伝子発現パターン表示方法および装置

【特許請求の範囲】

【請求項１】　時間経過に伴って発現の度合いが変化する複数の遺伝子の時系列発現パターンを視覚的に表示する遺伝子発現パターン表示方法であって、

前記複数の遺伝子の時系列発現パターンデータの任意の時間区間を指定するステップと、

指定された時間区間における時系列発現パターンデータを予め定めた基準値によってクラスタリングし、さらに同一クラスタ内において基準値を変えながら正または負の時間方向にクラスタリングを繰り返し行ない、その結果を予め定めた表示形式で表示するステップと

を備えることを特徴とする遺伝子発現パターン表示方法。

【請求項２】　前記基準値は、異なる遺伝子において発現のパターンが同じまたは異なるとみなすべき値であることを特徴とする請求項１記載の遺伝子発現パターン表示方法。

【請求項３】　前記時間区間において、異なる２つ以上の遺伝子が、初め同じ発現パターンを示し、途中から異なる発現パターンを示すものを予め定めた表示形式で表示することを特徴とする請求項１または２記載の遺伝子発現パターン表示方法。

【請求項４】　前記時間区間において、異なる２つ以上の遺伝子が、初め異なる発現パターンを示し、途中から同じ発現パターンを示すものを予め定めた表示形式で表示することを特徴とする請求項１または２記載の遺伝子発現パターン表示方法。

【請求項５】　時間経過に伴って発現の度合いが変化する複数の遺伝子の時系列発現パターンデータをデータベースから取得し、時系列発現パターンを表示装置画面に視覚的に表示する遺伝子発現パターン解析装置であって、

前記データベースから取得した前記複数の遺伝子の時系列発現パターンデータの任意の時間区間を指定する入力手段と、

指定された時間区間における時系列発現パターンデータを予め定めた基準値に

よってクラスタリングし、さらに同一クラスタ内において基準値を変えながら正または負の時間方向にクラスタリングを繰り返し行ない、その結果を予め定めた表示形式で前記表示装置画面に表示させる演算手段と
を備えることを特徴とする遺伝子発現パターン解析装置。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は、特定の遺伝子とハイブイリダイズさせることによって得られた時系列の遺伝子発現パターンを視覚的に分かり易く、そして遺伝子の機能・役割が推測し易い表示形式（または出力形式）によって表示するための表示方法および装置に関するものである。

【０００２】

【従来の技術】

ゲノム配列が決定された種の増加に伴い、進化に対応するとみられる遺伝子を見つけ出し、どの生物にも共通に持っていると考えられる遺伝子の集合を探したり、それから逆に種に個別な特徴を推測するなど、種間の遺伝子の違いから何かを見出そうとする、いわゆるゲノム比較法が盛んに行われてきた。

【０００３】

しかし近年、ＤＮＡチップやＤＮＡマイクロアレイなどのインフラストラクチャの発達によって、分子生物学の興味は、種間の情報から種内の情報へ、すなわち同時発生解析へと移りつつあり、これまでの種間の比較と合わせて、情報の抽出から関連付けの場が大きく広がりを持ち始めている。

【０００４】

例えば、既知の遺伝子と同一の発現パターンを示す未知の遺伝子が見つかれば、それが既知の遺伝子と同様の機能があると類推できる。これら遺伝子や蛋白質そのものの機能的な意味付けは、機能ユニットや機能グループといった形で研究されている。またそれらの間の相互作用も、既知の酵素反応データや物質代謝データとの対応付けによって、あるいはより直接的に、ある遺伝子を破壊あるいは過剰反応させ、その遺伝子の発現をなくすか、あるいは多量に発現させ、その遺

伝子の直接的および間接的影響を、全遺伝子の発現パターンを調べることによって解析している。

【０００５】

この分野において成功した事例として、スタンフォード大学のP.Brownらのグループによるイースト菌の発現解析が挙げられる（Michel B.Eisen et. al.: Cluster analysis and display of genome-wide expression patterns: Proc.Natl.Acad.Sci.(1998) Dec 8;95(25):14863-8）。彼らは、ＤＮＡマイクロアレイを用いて、細胞から抽出した遺伝子を時系列にハイブリダイズさせ、遺伝子の発現の度合い（ハイブリダイズした蛍光シグナルの輝度）を数値化した。数値に色を対応させることで、遺伝子の個々の発現過程を分かり易く表示させている。このとき、細胞の一連のサイクルにおいて発現パターンの過程が近い遺伝子同士（任意の時点での発現の度合いが近いもの同士）をクラスタリングしている。

【０００６】

図１３は、この手法に従って遺伝子の発現状態１３００を表示した例を示す図であり、横方向に時間軸、縦方向に遺伝子を並べている。このような表示方法をとることで、共通のクラスタに属する遺伝子は、共通の機能的性質をもつと類推することができる。なお、図１３における１つ１つの枠１３０１が１つの遺伝子のある時刻における発現状態を示すものであり、図１３では白黒の濃度を変えて発現状態を模式的に示している。

【０００７】

【発明が解決しようとする課題】

ところが、実際の遺伝子間の発現過程では、細胞の全サイクルにおいて同様の発現パターンを持つ幾つかの遺伝子グループを見つけ出すことで、その細胞全ての遺伝子間の関連が解明されるというほど単純ではない。

【０００８】

例えば、ある時点において異なる遺伝子が同じ機能のために同様に発現しているが、その後、次のある時点では別々の役割を持つような場合がある。当然この場合、遺伝子の発現過程は異なる。細胞の全サイクルにおいて発現のパターンが近いもの同士をクラスタリングして表示させる従来技術の手法では、これらの遺

伝子は別々のクラスタとして分類されるため、こういった性質を見つけ難いという難点があった。

【０００９】

本発明は、このような従来技術の問題点を鑑み、ある時点において異なる遺伝子が同じ機能のために同様に発現しているが、ある時点では別々の役割を持つような場合を見つけ出し、これを効果的に表示することが可能な遺伝子発現パターン表示方法および装置を提供することを目的とする。

【００１０】

【課題を解決するための手段】

本発明では、前記目的を達成するために、時間経過に伴って発現の度合いが変化する複数の遺伝子の時系列発現パターンを視覚的に表示する遺伝子発現パターン表示方法であって、

前記複数の遺伝子の時系列発現パターンデータの任意の時間区間を指定するステップと、指定された時間区間における時系列発現パターンデータを予め定めた基準値によってクラスタリングし、さらに同一クラスタ内において基準値を変えながら正または負の時間方向にクラスタリングを繰り返し行い、その結果を予め定めた表示形式で表示するステップとを備えることを特徴とする。

【００１１】

また、前記時間区間において、異なる２つ以上の遺伝子が、初め同じ発現パターンを示し、途中から異なる発現パターンを示すものを予め定めた表示形式で表示することを特徴とする。

【００１２】

また、前記時間区間において、異なる２つ以上の遺伝子が、初め異なる発現パターンを示し、途中から同じ発現パターンを示すものを予め定めた表示形式で表示することを特徴とする。

【００１３】

また、時間経過に伴って発現の度合いが変化する複数の遺伝子の時系列発現パターンデータをデータベースから取得し、時系列発現パターンを表示装置画面に視覚的に表示する遺伝子発現パターン解析装置であって、

前記データベースから取得した前記複数の遺伝子の時系列発現パターンデータの任意の時間区間を指定する入力手段と、指定された時間区間における時系列発現パターンデータを予め定めた基準値によってクラスタリングし、さらに同一クラスタ内において基準値を変えながら正または負の時間方向にクラスタリングを繰り返し行い、その結果を予め定めた表示形式で前記表示装置画面に表示させる演算手段とを備えることを特徴とする。

【００１４】

【発明の実施の形態】

以下、図面を参照して本発明の実施の形態を説明する。

図１は、本発明の遺伝子発現パターン表示方法を適用した遺伝子発現パターン解析装置の一実施形態を示すシステム構成図である。この実施形態の解析装置は、一連の細胞のプロセスにおいて遺伝子の発現の度合いを数値化した遺伝子発現パターンデータを格納した記憶装置（またはデータベース）１０１、発現パターンデータを視覚化して表示するための表示装置１０２、本システムへの値の入力や選択の操作を行なうためのキーボード１０３およびマウス１０４、遺伝子の発現過程に応じて発現パターンデータのクラスタリングを行なうクラスタリング処理部１０５から構成される。このクラスタリング処理部１０５は、コンピュータとそのプログラムによって具体化されるものである。

【００１５】

ここで、記憶装置１０１に代えて、ネットワーク等を介して遠隔地に設置されたサーバコンピュータが管理しているデータベースから遺伝子発現パターンデータを取得する構成にする実施形態がある。

【００１６】

本実施形態においては、細胞の一連のサイクルにおいて特定の時間区間を指定し、その時間区間において細かい粒度でクラスタリングを行なう。

【００１７】

すなわち、同一のクラスタに属する遺伝子は１つに束ね、異なるクラスタとの間には線を引き、さらに、クラスタ内の遺伝子において更にクラスタリングを行なう。細かい粒度のクラスタリングを範囲の始めから正の時間方向へ繰り返し行

なうと、図２に示すように、遺伝子の発現過程が木構造のように分岐して表現できる。図２において、２０１は、指定された時間区間、すなわちクラスタリング範囲である。

【００１８】

これは、指定された時間区間の始めにおいて同じ発現パターンを示し、時間区間の途中で異なる発現パターンを示したことを意味している。このような表示が得られた場合、始めの時点では異なる遺伝子が同じ機能のために同様に発現しているが、ある時点において別々の役割を持つため異なって発現したと類推することができる。

【００１９】

同様に、細かい粒度のクラスタリングを範囲の終端から負の時間方向へ繰り返し行なうと、遺伝子の発現の過程が、図３のように、逆の木構造のような分岐構造として表現することができる。

【００２０】

これは、範囲の始めにおいて異なる発現パターンを示し、範囲の途中で同じ発現パターンを示したことを意味している。このような表示が得られた場合、始めの時点では異なる遺伝子が異なる機能を持っていたが、ある時点において同様の役割を持ったと類推することができる。

【００２１】

図４は、遺伝子の発現パターンデータをクラスタリングして表示するクラスタリング処理部１０５におけるアルゴリズムの概要を示すフローチャートである。

【００２２】

ここではまず、初期パラメータを設定し（ステップ４０１）、表示位置決定処理を行なう（ステップ４０２）。初期パラメータについては、後述する。その後、表示処理を行ない、処理を終了する（ステップ４０３）。本アルゴリズムは、図２に示したように、異なる遺伝子が、ある時間区間において、始めにおいて同じ発現パターンを示し、途中で異なる発現パターンを示したことを表示するものである。

【００２３】

図５は、本アルゴリズムで使われる変数と実データとの対応関係を示す説明図である。図６は、図４中の初期パラメータ設定処理（ステップ４０１）に関するアルゴリズムの詳細を示している。

【００２４】

まず、遺伝子発現パターンデータを記憶装置１０１から読み込む。この遺伝子発現パターンデータには、図５に示すようにｍ＋１個のサンプル遺伝子$g_0, g_1, \ldots g_m$について、時刻$T_0, T_1, \ldots T_n$において実験した結果の発現パターンデータが入っているものとする。そこで、時刻T_jにおける遺伝子g_iの発現の観測値をg[j][i]とおく（ステップ６０１）。

【００２５】

次に、キーボード１０３、マウス１０４を使って、クラスタリング適用範囲（開始時刻T_{start}、終了時刻T_{end}）、異なるクラスタとみなすべき基準を示す正数値（$K_{start}, K_{start+1}, \cdots, K_{end}$）、クラスタリングの粒度を示す整数（Ｓ）、クラスタリング手法をそれぞれ入力する（ステップ６０２）。

【００２６】

クラスタリング適用範囲とは、図２、図３に太枠実線２０１で示すように、細胞の一連のプロセスにおいて、より詳しくクラスタリングする時間区間を示す。例えば細胞の一連のプロセスにおいて、ある時刻で細胞に特殊な発現パターンがみられた場合、その時刻の前後をクラスタリング適用範囲に指定することで、全遺伝子の発現状態をより詳しくモニタリングするように選択する。従来のクラスタリングとの基本的な相違点は、図１３のような細胞の全プロセスにおいて発現状態の近いもの同士をクラスタリングするのではなく、図２に示すような相異なる遺伝子が範囲の始めにおいて同じ発現パターンを示し、範囲の途中で異なる発現パターンを示したことを表示するところにある。

【００２７】

異なるクラスタとみなす基準とは、異なるクラスタの間の非類似度が最低でもどれくらいの値をとるかを示すものである。すなわち、クラスタ間の閾値Ｋを示している。閾値が$K_{start}, K_{start+1}, \cdots, K_{end}$と可変に設定できることで、時間

によって粗いクラスタリングから細かいクラスタリングまで調節できる。

【００２８】

また、クラスタリングを行なうときの非類似度の計算において、本システムでは、発現データの時刻$T_0, T_1, \ldots T_n$における全てのデータを非類似度の計算の対象とせずに、ある時間区間を設けて、その時間区間内におけるデータを非類似度の計算の対象とする。この時間区間を図５に示すようにスリット５０１、このスリット５０１の長さ（時間軸方向の幅）Ｓをクラスタリングの粒度とよぶ。本アルゴリズムでは、まずスリット５０１の先頭をT_{start}に合わせてデータをT_{start}から$T_{start+S}$の範囲でクラスタリングを行ない、そこで分割された各々のクラスタ内において、スリット５０１を時刻が正の方向へ１つずらし、$T_{start+1}$から$T_{start+S+1}$の範囲でクラスタリングを行なう。このような操作をスリットの後端がT_{end}になるまで逐次実行する。したがって、粒度が細かいほど、すなわち時間区間の幅が短いほど、より細かい遺伝子間の発現の違いを表すことができる。

【００２９】

クラスタリング手法では、クラスタリングにおいて個体同士の相関関係を表す類似度または非類似度（ピアソンの相関係数、ユークリッド平方距離、標準化ユークリッド平方距離、マハラノビスの距離、ミンコフスキー距離など）及びクラスタ合併のアルゴリズム（最短距離法、最長距離法、群平均法、重心法、メディアン法、ウォード法、可変法など）を指定する。本アルゴリズムは非類似度を対象としているが、クラスタリング手法において類似度を選択した場合は、計算した類似度に負符号を付けたり、逆数をとるなどの操作を施し、非類似度に変換すればよい。

【００３０】

これらの値を設定したら、それぞれの項目が正しいかどうか調べる。クラスタリング適用範囲T_{start}、T_{end}がT_0からT_nの範囲に含まれているか（ステップ６０３）、クラスタリングの粒度Ｓがクラスタリング適用範囲の幅を超えてないか（$S \leqq end-start$）（ステップ６０４）、また設定したクラスタリング手法において、合併アルゴリズムを重心法、メディアン法、ウォード法を選択した時

、非類似度においてユークリッド平方距離を選択したかなど、類似度または非類似度と合併アルゴリズムは妥当な組み合わせか（ステップ６０６）を調べる。もし、これらの値で正当なものが入っていないならば、表示装置１０２にエラーを出力し、再入力を促す（ステップ６０７）。

【００３１】

しかし、設定項目が適切であった場合、次に、i=1,2,…,mに対して遺伝子g_iの平均発現度G_i=（g[0][i]+g[1][i]+…g[n][i]）/nを求める（ステップ６０８）。

【００３２】

次に、個々の遺伝子の表示情報を格納するために図５に示すような配列l[I](I=0,1,…,m)５０２と整数値変数lmaxを用意する。各l[I]は構造体データで、図５に示すように遺伝子のインデックスを表すメンバ（index）と異なるクラスタ間の仕切り線の位置を表すメンバ(linepos)からなる。構造体のメンバは、l[I].index,l[I].linesposという形で値を代入・参照できる。そこで、全てのIに対してl[I].lineposの値をT_{end}として初期化し（ステップ６０９）、さらにlmaxの値を「０」としておく（ステップ６１０）。次に、変数ｔにstartの値を設定する（ステップ６１１）。

【００３３】

本アルゴリズムでは、整数値の集合を表す“クラスタ”と呼ばれる抽象データ型を使っている。クラスタには、整数の登録、削除、登録データの参照のインタフェースを備えているものとする。

【００３４】

クラスタＢを生成し、そこに{0,1,2,…,m}を登録し処理を終了する（ステップ６１２）。

【００３５】

以上のように初期設定をした後、クラスタリング適用範囲２０１に対して処理を行なう。すなわち、上で定めたｔとＢとを引数として表示位置決定処理（図４のステップ４０２の処理Ａ）を行なう。

【００３６】

図７は、図４中の表示位置決定処理（処理Ａ）の詳細を示すフローチャートであり、この処理Ａの中で配列lに表示情報を登録する。

【００３７】

まず、引数として渡されたクラスタをＢ、時刻をｔとする（ステップ７０１）。ここでＢを更にクラスタリングする（処理Ｂ）。このときｔとＢを引数として与える。処理Ｂの結果として、総クラスタ数がcmaxに、クラスタリング結果がＡ［Ｊ］（Ｊ=1,2,…,cmax)に設定される（ステップ７０２）。処理Ｂの詳細については後述する。

【００３８】

次に、「ｔ＋Ｓ」がendと等しいかどうか判定する（ステップ７０３）。endの時はスリット５０１の終端がクラスタリング適用範囲２０１の終わりに来たことを意味し、ここでクラスタリング処理を終了する。このとき、Ｊ=１としてＪがcmaxを超えるまで、各々のクラスタに対して次の処理を実行する（ステップ７０４，７０５）。クラスタＡ[Ｊ]の要素が $\{i_1,\cdots,i_k\}$ であるとき、これらの要素を一定の基準の下に並べて表示する。ここでは各要素に対応する遺伝子の平均発現度 $G_{i1},\ldots G_{ik}$ を値の降順に並べて、それを $G_{j1},\ldots G_{jk}$ とおく（ステップ７０６）。

【００３９】

次に配列lの値を入力する。すなわち、発現パターンデータの位置情報を表すl[].indexに平均輝度が降順になるように l[lmax].index=j_1、l[lmax+1].index=j_2、…、l[lmax+k−1].index=j_kと設定し（ステップ７０７）、異なるクラスタとの仕切り線（図２の２０２で代表して示す横方向の太実線）を表す l[lmax+k−1].linepos に時刻ｔからｔ+Ｓ（＝T_{end}）の範囲まで線を引くことを示すｔの値を入力する（ステップ７０８）。

【００４０】

次に、配列lの入力済みデータの最大数を示すlmaxにkを加算する（ステップ７０９）。次に、Ｊを１つインクリメントし、次のクラスタの処理に移る（ステップ７１０）。

【００４１】

一方、ステップ７０３において、「ｔ+Ｓ」がendと一致しない場合、すなわちスリット５０１の終端がクラスタリング適用範囲２０１の終わりに来ていないとき、ｔを１つインクリメントし、Ｊに「１」を設定する（ステップ７１１）。Ｊがcmaxを超えるまで、各々のクラスタに対して次の処理を行なう（ステップ７１２）。すなわちＢにＡ[Ｊ]を代入し（ステップ７１３）、引数として時刻ｔ、クラスタＢを与えて表示位置決定処理（処理Ａ）を行なう（ステップ７１４）。次に、異なるクラスタとの仕切り線を表す l[lmax−1].linepos に時刻ｔからT_{end}の範囲まで線を引くことを示すｔを入力する（ステップ７１５）。そして、Ｊを１つインクリメントし、次のクラスタの処理に移る（ステップ７１６）。全てのクラスタＡ[Ｊ]（Ｊ=1,…,cmax）に関する処理が終われば終了する。

【００４２】

図８および図９は、クラスタリング処理（処理Ｂ）のアルゴリズムを示すフローチャートである。

まず、引数として入力されたクラスタをＢ、入力された時刻をｔに入れる（ステップ８０１）。

【００４３】

次に、クラスタBの要素が$i_1,\cdots,i_k$であるとき、$i_1,\cdots,i_k$に対応する遺伝子間の時刻ｔから時刻ｔ+Ｓにおける類似度または非類似度d_{ij}($i<j$かつ$i,j\in\{i_1,i_2,\cdots,i_k\}$)を求める（ステップ８０２）。

【００４４】

ここで、遺伝子g_i,g_jに対する遺伝子発現データ {g[0][i],g[1][i],…,g[n][i]}、{g[0][j],g[1][j],…,g[n][j]} の時刻ｔから時刻ｔ+Ｓにおける類似度（非類似度）とは、例えば以下のような計算で求める量である（ステップ８０２）。

【００４５】

（１）類似度としてピアソンの相関係数を指定したとき

【００４６】

【数１】

$$d_{i,j} = \frac{\sum_{k=i}^{i+S}(g[k][i]-\overline{g[i]})(g[k][j]-\overline{g[j]})}{\sqrt{\left\{\sum_{k=i}^{i+S}(g[k][i]-\overline{g[i]})^2\right\}\left\{\sum_{k=i}^{i+S}(g[k][j]-\overline{g[j]})^2\right\}}} \quad \cdots\cdots (1)$$

ただし $\overline{g[l]} = \frac{1}{S}\sum_{k=i}^{i+S} g[k][l]$

【００４７】

となる。本アルゴリズムでは非類似度を対象にしているので、類似度を適用する場合には負符号を付ける、逆数をとるなどの操作をして非類似度に変換しなければならない。

【００４８】

（２）非類似度としてユークリッド平方距離を指定したとき、

【００４９】

【数２】

$$d_{i,j} = \sum_{k=i}^{i+S}(g[k][i]-g[k][j])^2 \quad \cdots\cdots (2)$$

【００５０】

（３）標準化ユークリッド平方距離を指定したとき、

【００５１】

【数３】

$$d_{i,j} = \sum_{k=i}^{i+S}(g[k][i]-g[k][j])^2 / s_k^2 \quad \cdots\cdots (3)$$

ただし s_k^2 は変量g[k][0],...,g[k][n]の分散。

【００５２】

（４）マハラノビスの距離を指定したとき、

【００５３】

【数４】

$$d_{i,j}={}^t(g[i]-g[j])S^{-1}(g[i]-g[j]) \qquad \cdots\cdots (4)$$

ただし g[l]=^{t}(g[t][l],...,g[t+S][l])、
Sはg[i],g[j]の共分散行列。

【００５４】

（５）ミンコフスキー距離を指定したとき、

【００５５】

【数５】

$$d_{i,j}=\left\{\sum_{l=t}^{t+S}|g[l][i]-g[l][j]|^k\right\}^{1/k} \qquad \cdots\cdots (5)$$

【００５６】

クラスタC[1],…,C[k]を生成し、それぞれのクラスタにC[1]←{i_1},……,C[k]←{i_k}を登録しておく（ステップ８０３）。そして、生成したクラスタの数を表す変数ccntにkを代入しておく（ステップ８０４）。次に、空集合のクラスタDを生成する（ステップ８０５）。

【００５７】

次に、ここまでで計算した非類似度$d_{i,j}$(i,j ∈{1,2,…,ccnt}－D)の値の最小値$d_{p,q}$を求め、先に設定した閾値K_t以下かどうか判定する（ステップ８０６、８０７）。$d_{p,q}$がK_t以下のとき次のことを実行する。クラスタC[ccnt+1]を新たに生成し、クラスタC[p]とクラスタC[q]に含まれる要素の和集合をクラスタC[ccnt+1]に登録し（ステップ８０８）、クラスタC[p]とクラスタC[q]に含まれる要素を削除する（ステップ８０９）。次に、C[p]とC[q]はもう必要ないので、Dにp、qを登録する（ステップ８１０）。次に、クラスタC[h]（h ∈{1,2,…,ccnt}－D)とクラスタC[ccnt+1]間の時刻ｔから時刻「ｔ+Ｓ」における非類似度 $d_{h,ccnt+1}$を求める（ステップ８１１）。ここで$d_{h,ccnt+1}$は、次の計算式で求めることができる。すなわち

【００５８】

【数６】

$$d_{k,ccnt+1} = \alpha d_{k,p} + \beta d_{k,q} + \gamma d_{p,q} + \delta \left| d_{k,p} - d_{k,q} \right| \quad \cdots\cdots (6)$$

【００５９】

ここでα、β、γ、δは、n(k)をクラスタＣ[k]内の要素の個数としたとき、クラスタリング手法が

（１）最短距離法のときα=0.5、β=0.5、γ=0、δ=－0.5

（２）最長距離法のときα=0.5、β=0.5、γ=0、δ=0.5

（３）群平均法のときα=n(p)/n(ccnt+1)、β=n(q)/n(ccnt+1)、γ=0、δ=0

（４）重心法のときα=n(p)/n(ccnt+1)、β=n(q)/n(ccnt+1)、

$\gamma = -n(p)n(q)/n(ccnt+1)^2$、δ=0

（５）メディアン法のときα=0.5、β=0.5、γ=－0.25、δ=0

（６）ウォード法のときα={n(h)+n(p)}/{n(h)+n(ccnt+1)}、

β={n(h)+n(q)}/{n(h)+n(ccnt+1)}、γ=－n(h)/{n(h)+n(ccnt+1)}、δ=0

である。

【００６０】

次に、生成したクラスタの数を表す変数ccntに「１」を加える（ステップ８１２）。これらの処理を更新した$d_{i,j}$(i,j∈{1,2,…,ccnt}－D)の最小値がK_tより大きくなるまで続ける。

【００６１】

ステップ８０７において$d_{i,j}$の最小値$d_{p,q}$がK_tより大きいとき、クラスタリングを終えて、結果の出力処理を行なう。まず、クラスタＣ[1]からＣ[ccnt]で、空集合でないものを判定し、この総数をcmaxに入力する（ステップ８１３）。そして、cmax個のクラスタＡ[1],…,Ａ[cmax]を生成する（ステップ８１４）。空集合でないクラスタに対し、それに含まれる遺伝子の平均発現度の平均をとる。すなわち、クラスタＣ[p]={i_1,…,i_k}に対して、$G'_p = (G_{i1}+\ldots+G_{ik})/k$を求める。この値を降順に並べたものを、$G'_{p1}$,…,$G'_{pcmax}$としたときＡ[1] ← Ｃ[$p_1$],…,Ａ[cmax] ← Ｃ[$p_{cmax}$]を登録する（ステップ８１５）。最後に

、総クラスタ数cmaxとクラスタＡ[1],…,Ａ[cmax]を出力し（ステップ８１６）、処理を終了する。

【００６２】

図１０は、図４における表示処理のアルゴリズムの詳細を示すフローチャートである。このアルゴリズムは、配列l[I]を読み込み、対応する遺伝子の発現データを表示する処理である。

【００６３】

まずiの値を「０」とし（ステップ１０００）、iの値がlmaxと等しくなるまで、各々の遺伝子発現データに対して以下の操作を続ける（ステップ１００１）。次に、x=l[i].indexが指す遺伝子１行分の発現データg[k][x](k=0,1,…,n)の数値を対応する表示色に置き換え、第i行として１行にわたり表示する（ステップ１００２）。更に、クラスタ間の仕切り線を、今表示した第i行のすぐ下の時刻l[i].linepos からT_{end}の範囲に引く（ステップ１００３）。

【００６４】

ここで、l[i].linesposの値が、初期値T_{end}の場合は、クラスタ間の仕切り線は存在せず線も書く必要が無い。iを１つずつインクリメントし（ステップ１００４）、ステップ１００１においてiがlmaxになったら、処理を終える。

【００６５】

以上の処理によって、図２に示したような、相異なる遺伝子がクラスタリング適用範囲の始めにおいて同じ遺伝子発現パターンを示し、範囲の途中で異なる発現パターンを示すような状況を効果的に表示することができる。

【００６６】

また、図３に示したような、相異なる遺伝子がクラスタリング適用範囲の始めにおいて異なる遺伝子発現パターンを示し、範囲の途中で同じ発現パターンを示すような状況を効果的に表示する場合には、ステップ６０９（図６）においてl[i].linesposにT_{start}を、ステップ６１１においてｔにendを設定し、ステップ７０３（図７）においてｔ+Ｓ=endの判定条件をｔ－Ｓ=startにし、ステップ７１１においてｔ←ｔ+1をｔ←ｔ－1に置き換え、ステップ１００３（図１０）においてクラスタ間の仕切り線を、T_{start}からl[i].linesposの範囲に引けばよい。

これは、はじめスリットの終端部分をT_{end}に設定しておき、時間軸の負の方向へ１つずつスリットを移動してクラスタリングすることを意味している。

【００６７】

また、これらの詳細なクラスタリング手法の応用例として、クラスタリング適用範囲の前方から時間軸の正の方向へスリットを動かしてクラスタリングを行ない、図１１に示したような表示が得られた場合を考える。このとき、図１１の点線１１０１，１１０２で囲んだような似通った発現パターンが見られた場合、それらの遺伝子をマーキング（１１０３）しておき、クラスタリング適用範囲２０１の後方から時間軸の負の方向に向けてクラスタリングを行なう。もし、図１２に示したようにマーキング（１１０３）した遺伝子が互いに近い位置にあるものが見つかる（例えば①と④、③と⑥など）ならば、これらの遺伝子は始め異なる遺伝子発現パターンを示し、途中で同じ発現パターンを示すことを意味しており、このような双方向のクラスタリングによって個々の遺伝子の発現状態を容易に推測することが出来る。

【００６８】

更に、T_{start}をT_0にT_{end}をT_nに、スリット幅Ｓをnに設定すれば、従来の技術の中で説明したP.Brownらの結果と同様の表示を得ることが出来る。

【００６９】

なお、本発明は、上記実施形態に限定されるものではなく、実施に際しては、細部を種々変更して実施することができる。例えば、途中から発現パターンが変わった部分あるいは境界においては、フリッカ表示、高輝度表示、色反転表示などの既知の表示形態を各種組み合わせて表示することができる。

【００７０】

また、クラスタリング処理部１０５の処理は、プログラムとしてＣＤ－ＲＯＭ等の記録媒体に記録してコンピュータユーザに提供することができる。

【００７１】

また、遺伝子のデータとしては、時系列の発現データに限定されるものではなく、図３または図４における横軸（時間軸）を他の基準にとり変えることによって、例えば異なる実験間について比較を行うなどの利用が考えられる。

【００７２】

また、解析結果を表示装置画面に表示する例を説明したが、最近においては多色プリンタの精度が向上しているため、多色プリンタで印刷出力する構成であってもよい。本発明の表示とは、プリンタで視覚的に印刷出力する概念を含むものである。

【００７３】

【発明の効果】

以上説明したように、本発明によれば、細胞の発現サイクルの一部区間を指定し、その範囲において細かい粒度でクラスタリングを行なうことができる。そして、この表示結果に基づいて、利用者は遺伝子の発現経過の状態をより詳細に観測することができ、遺伝子の発現状態から生物学的機能を効率的よく推測することができる。

【図面の簡単な説明】

【図１】

本発明を適用した解析装置の一実施形態を示すシステム構成図である。

【図２】

クラスタリングの範囲を制限して細かい粒度でクラスタリングしたときの遺伝子発現パターン表示例（その１）を示す模式図である。

【図３】

クラスタリングの範囲を制限して細かい粒度でクラスタリングしたときの遺伝子発現パターン表示例（その２）を示す模式図である。

【図４】

クラスタリング処理の概要を示すフローチャートである。

【図５】

クラスタリング処理で使用する変数と実データの関係を示す説明図である。

【図６】

初期パラメータの設定に関するアルゴリズムを示すフローチャートである。

【図７】

表示位置決定処理のアルゴリズムを示すフローチャートである。

【図８】

クラスタリングのアルゴリズムを示すフローチャートである。

【図９】

図８の続きを示すフローチャートである。

【図１０】

表示処理のアルゴリズムの概要を示すフローチャートである。

【図１１】

クラスタリング適用範囲の前方から時間軸の正の方向へスリットを動かしてクラスタリングを行ったときの遺伝子発現パターン表示例を示す説明図である。

【図１２】

クラスタリング適用範囲の後方から時間軸の負の方向へスリットを動かしてクラスタリングを行ったときの遺伝子発現パターン表示例を示す説明図である。

【図１３】

細胞の全プロセスにおいて発現状態の近いものどうしをクラスタリングしたときの遺伝子発現パターン表示例を示す説明図である。

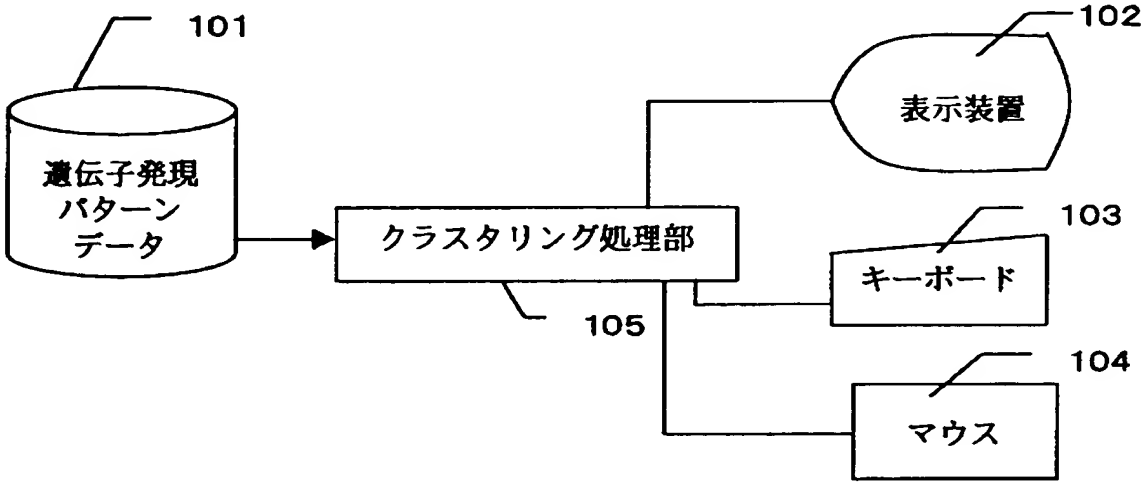
【符号の説明】

１０１…遺伝子発現パターンデータの記憶装置、１０２…表示装置、１０３…キーボード、１０４…マウス、１０５…クラスタリング処理部、２０１…クラスタリング範囲、５０１…スリット。

【書類名】　　図面

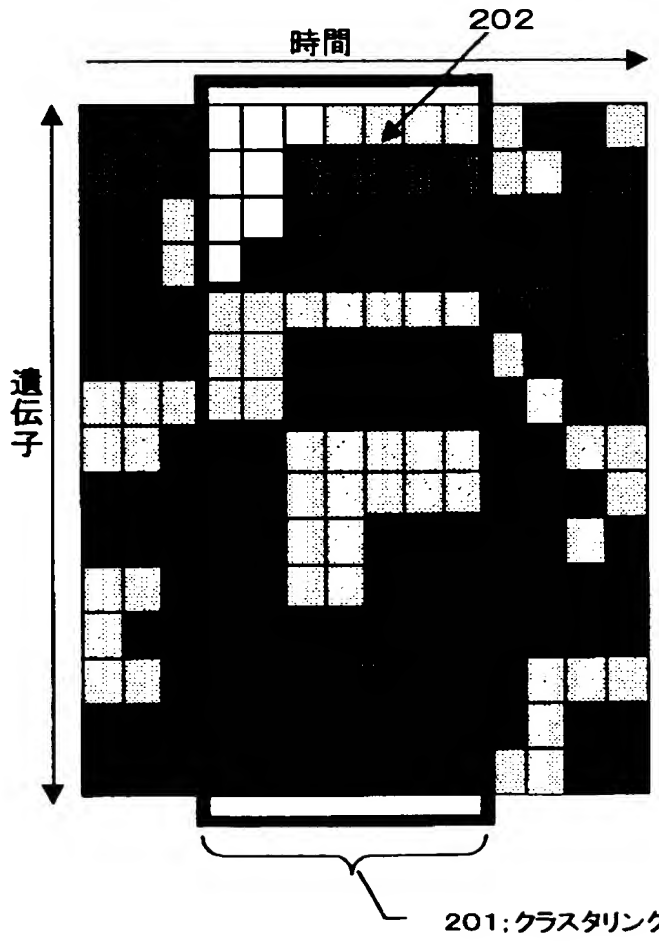
【図１】

図1



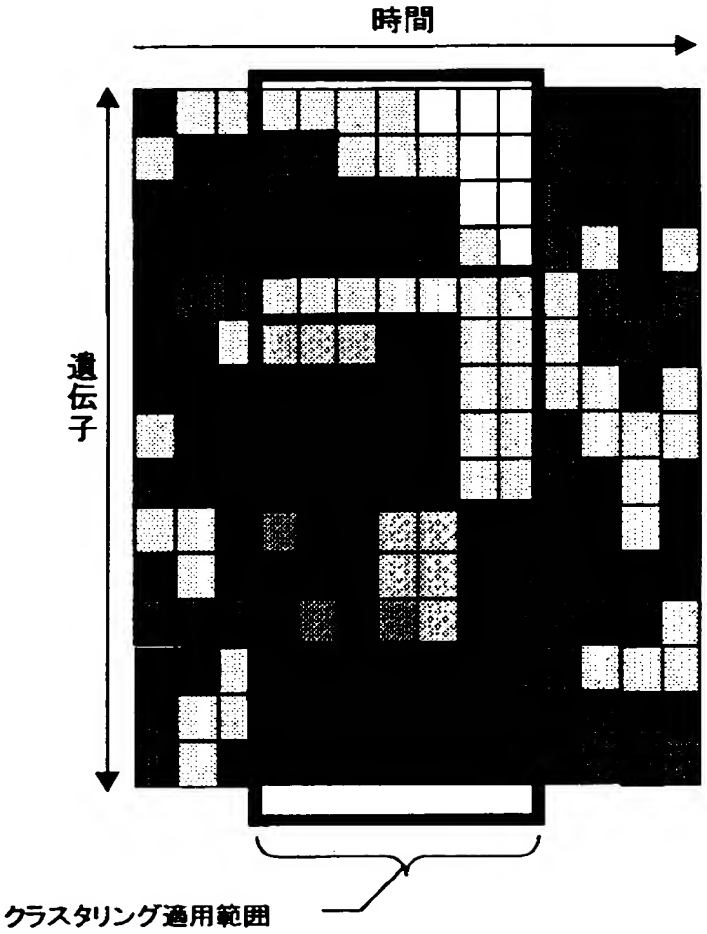
【図２】

図2



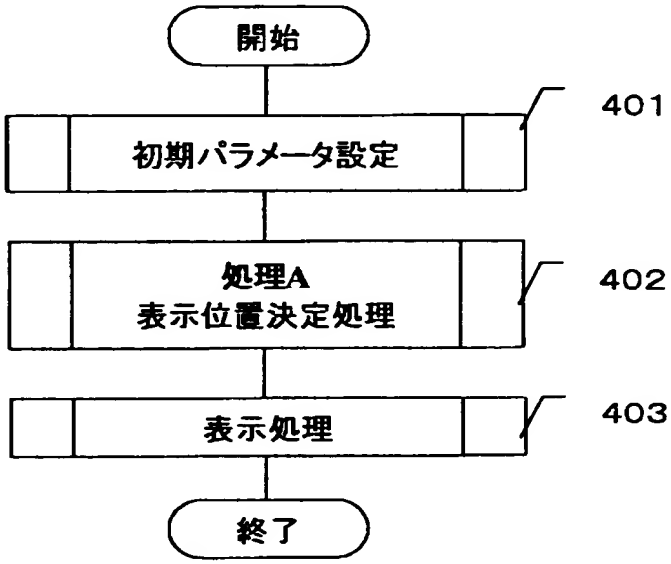
【図３】

図3



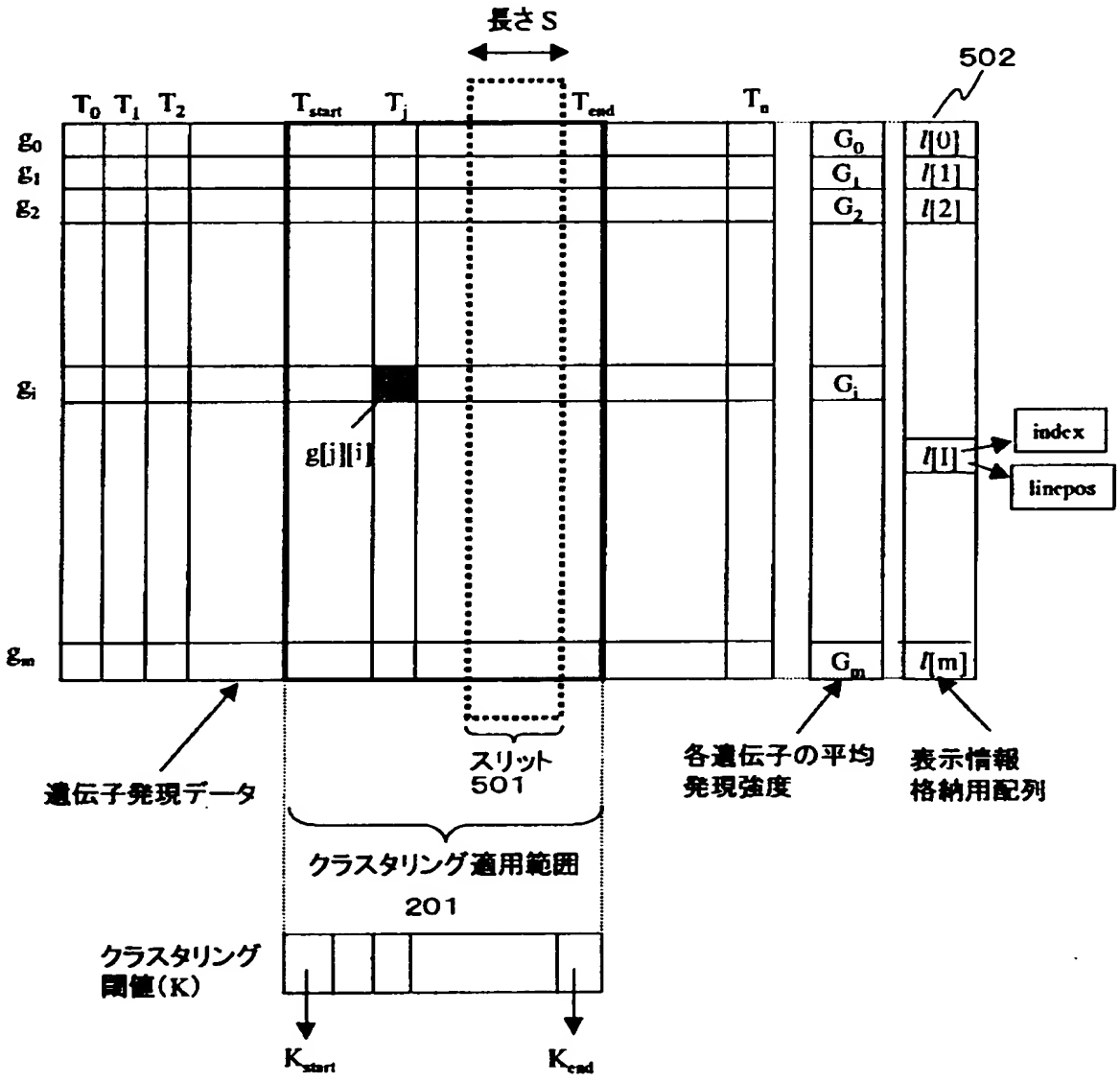
【図４】

図4



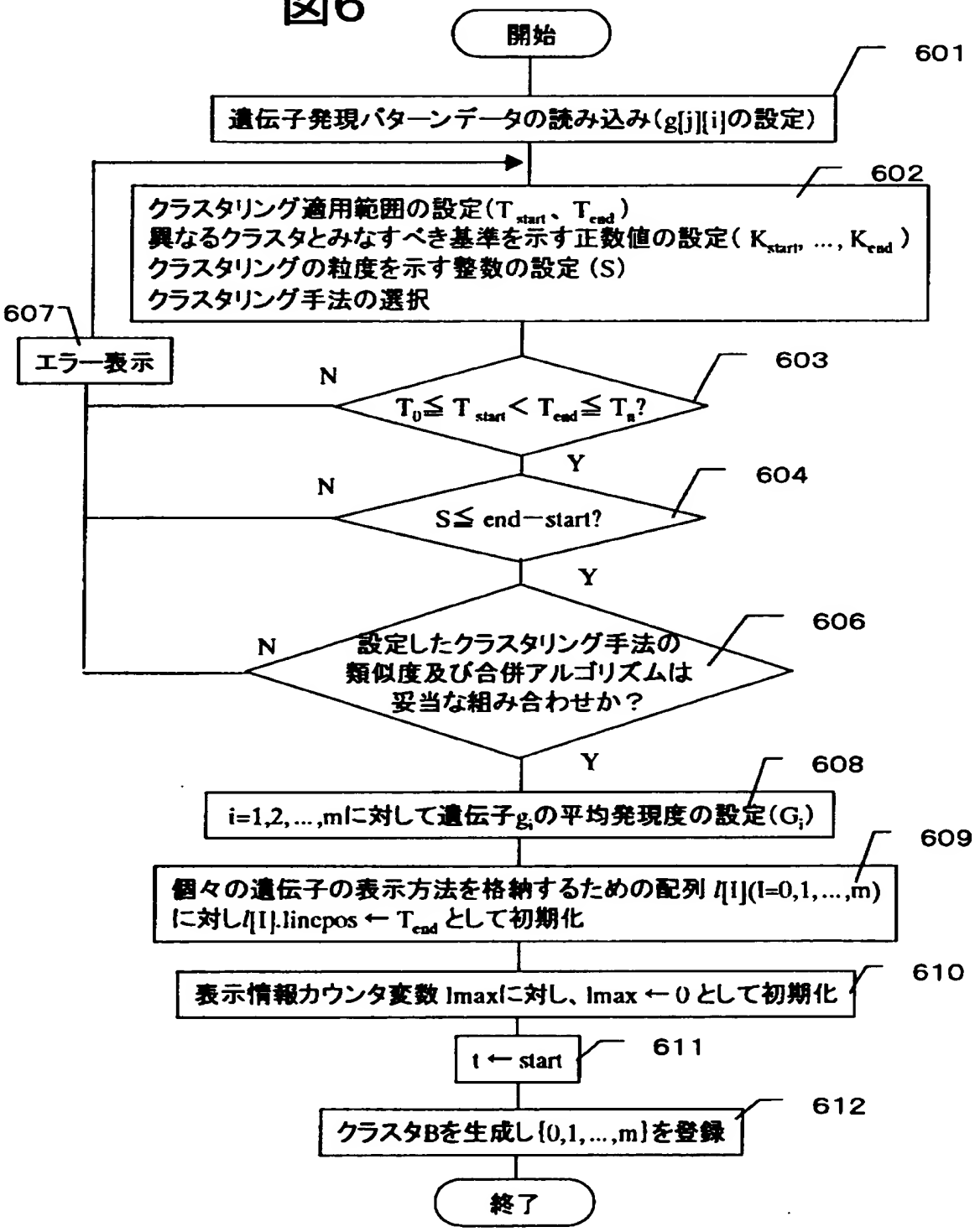
【図５】

図5


長さS
502
T0 T1 T2
Tstart
Tj
Tend
Tn
g0
g1
g2
gi
gm
G0
G1
G2
Gi
Gm
l[0]
l[1]
l[2]
l[I]
l[m]
g[j][i]
index
linepos
遺伝子発現データ
スリット
501
各遺伝子の平均
発現強度
表示情報
格納用配列
クラスタリング適用範囲
201
クラスタリング
閾値(K)
Kstart
Kend

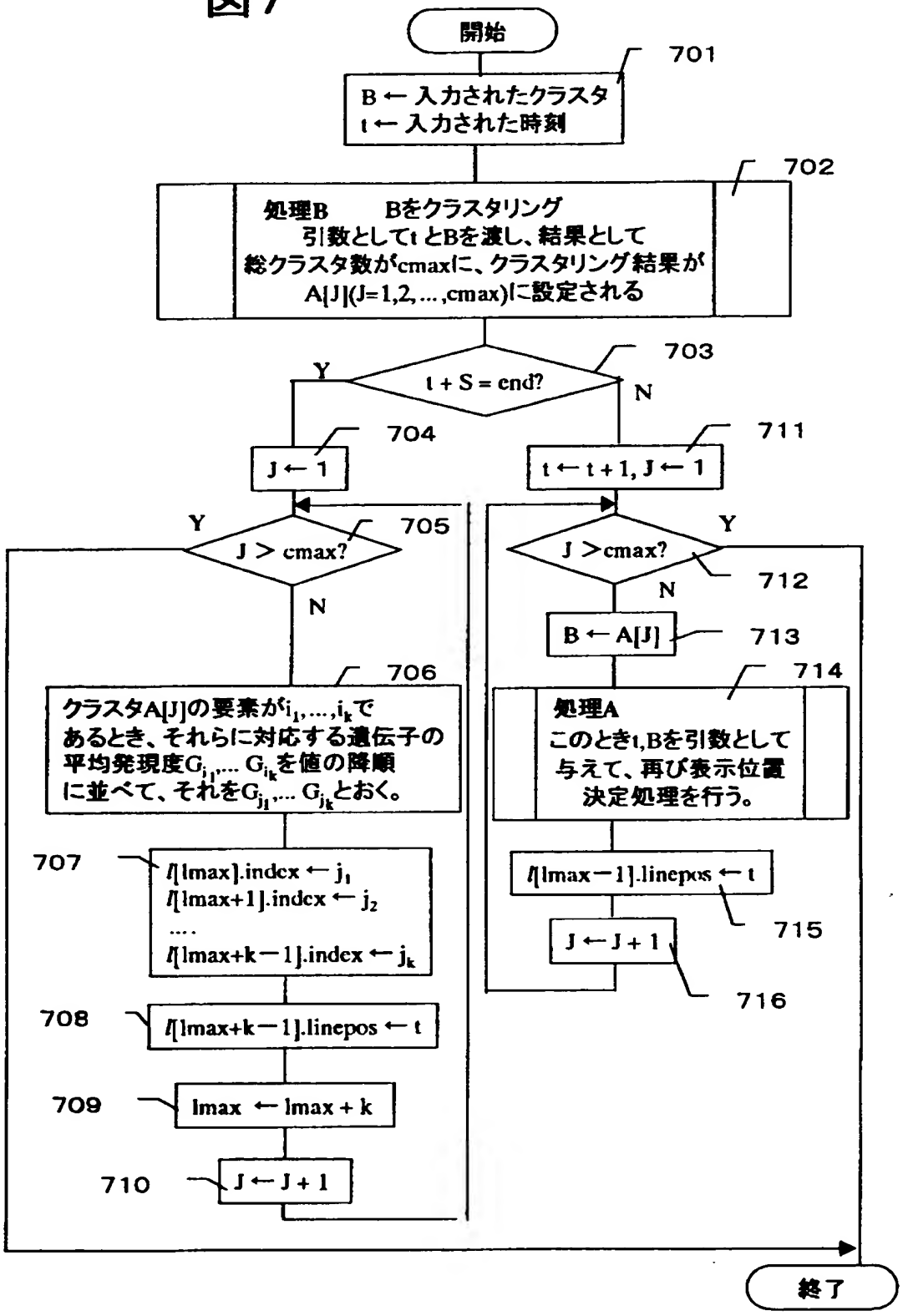
【図６】

図6


開始
601
遺伝子発現パターンデータの読み込み(g[j][i]の設定)
602
クラスタリング適用範囲の設定(T_{start}、T_{end})
異なるクラスタとみなすべき基準を示す正数値の設定(K_{start}, …, K_{end})
クラスタリングの粒度を示す整数の設定 (S)
クラスタリング手法の選択
607
エラー表示
603
N
$T_0 \leqq T_{start} < T_{end} \leqq T_n$?
Y
604
N
S≦ end−start?
Y
606
N
設定したクラスタリング手法の
類似度及び合併アルゴリズムは
妥当な組み合わせか？
Y
608
i=1,2,…,mに対して遺伝子g_iの平均発現度の設定(G_i)
609
個々の遺伝子の表示方法を格納するための配列 I[I](I=0,1,…,m)
に対しI[I].lincpos ← T_{end} として初期化
610
表示情報カウンタ変数 Imaxに対し、Imax ← 0 として初期化
611
t ← start
612
クラスタBを生成し{0,1,…,m}を登録
終了

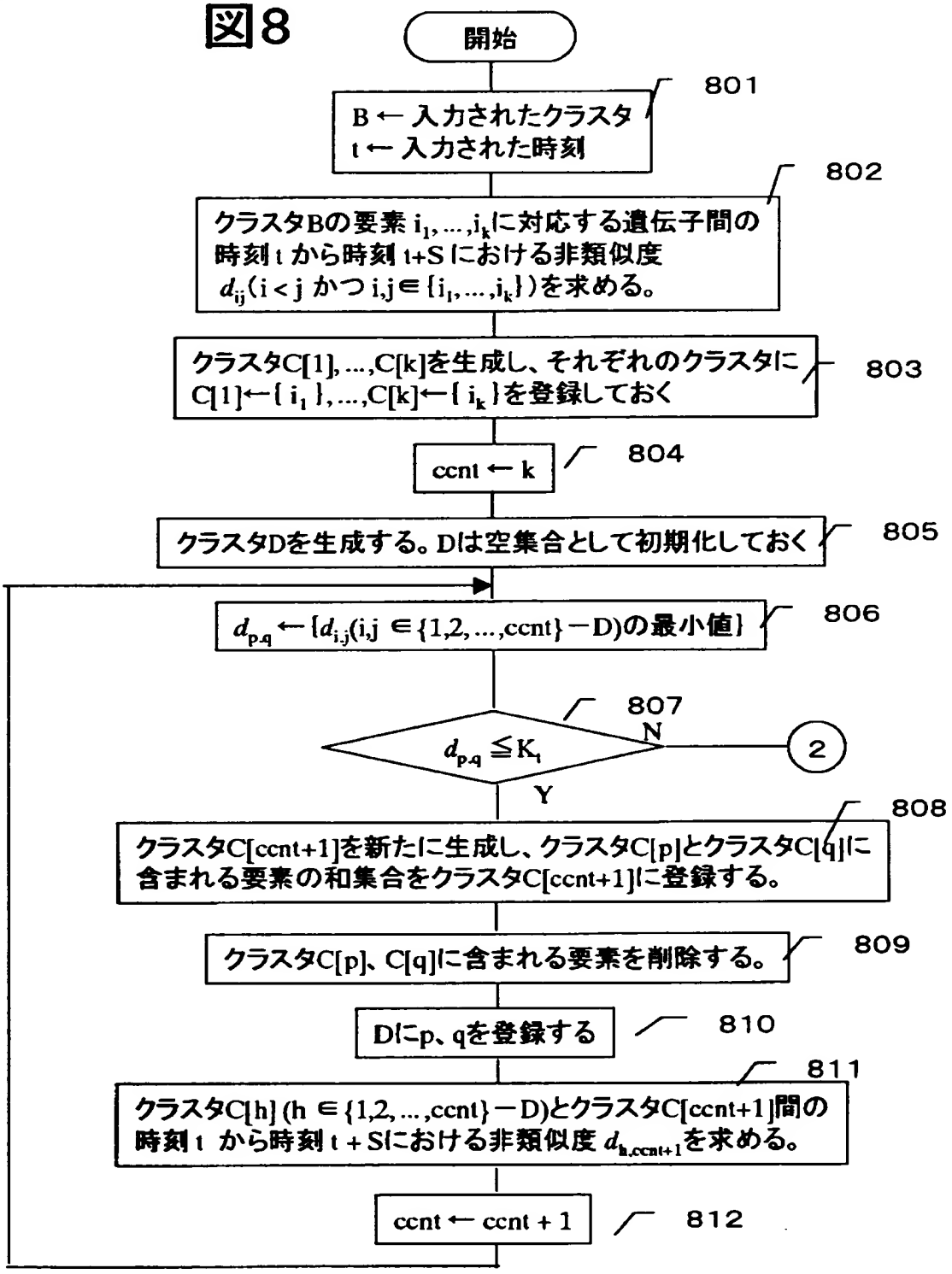
【図７】

図7


開始
701
B ← 入力されたクラスタ
t ← 入力された時刻
702
処理B　Bをクラスタリング
引数としてt とBを渡し、結果として
総クラスタ数がcmaxに、クラスタリング結果が
A[J](J=1,2,...,cmax)に設定される
703
t + S = end?
Y
N
704
J ← 1
711
t ← t + 1, J ← 1
Y
705
J > cmax?
N
706
クラスタA[J]の要素がi_1,...,i_kで
あるとき、それらに対応する遺伝子の
平均発現度G_{i_1},... G_{i_k}を値の降順
に並べて、それをG_{j_1},... G_{j_k}とおく。
707
I[lmax].index ← j_1
I[lmax+1].index ← j_2
....
I[lmax+k−1].index ← j_k
708
I[lmax+k−1].linepos ← t
709
lmax ← lmax + k
710
J ← J + 1
Y
J > cmax?
712
N
B ← A[J]
713
714
処理A
このときt,Bを引数として
与えて、再び表示位置
決定処理を行う。
I[lmax−1].linepos ← t
715
J ← J + 1
716
終了

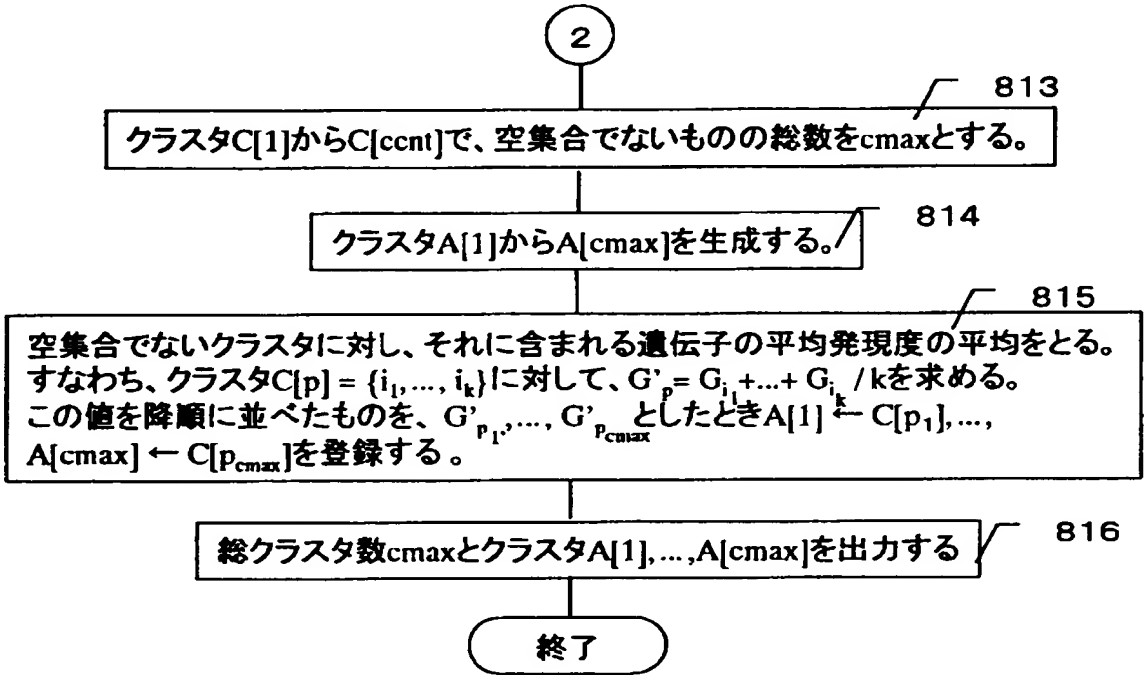
【図８】

図8


開始
801
B ← 入力されたクラスタ
t ← 入力された時刻
802
クラスタBの要素 $i_1,\ldots,i_k$ に対応する遺伝子間の時刻 t から時刻 t+S における非類似度 d_{ij} (i < j かつ $i,j \in \{i_1,\ldots,i_k\}$) を求める。
803
クラスタC[1],...,C[k]を生成し、それぞれのクラスタに C[1]←{ i_1 },...,C[k]←{ i_k }を登録しておく
804
ccnt ← k
805
クラスタDを生成する。Dは空集合として初期化しておく
806
$d_{p,q} \leftarrow \{d_{i,j}(i,j \in \{1,2,\ldots,ccnt\} - D)$の最小値$\}$
807
$d_{p,q} \leqq K_t$
N
2
Y
808
クラスタC[ccnt+1]を新たに生成し、クラスタC[p]とクラスタC[q]に含まれる要素の和集合をクラスタC[ccnt+1]に登録する。
809
クラスタC[p]、C[q]に含まれる要素を削除する。
810
Dにp、qを登録する
811
クラスタC[h] (h ∈ {1,2,...,ccnt} − D)とクラスタC[ccnt+1]間の時刻 t から時刻 t + Sにおける非類似度 $d_{h,ccnt+1}$ を求める。
812
ccnt ← ccnt + 1

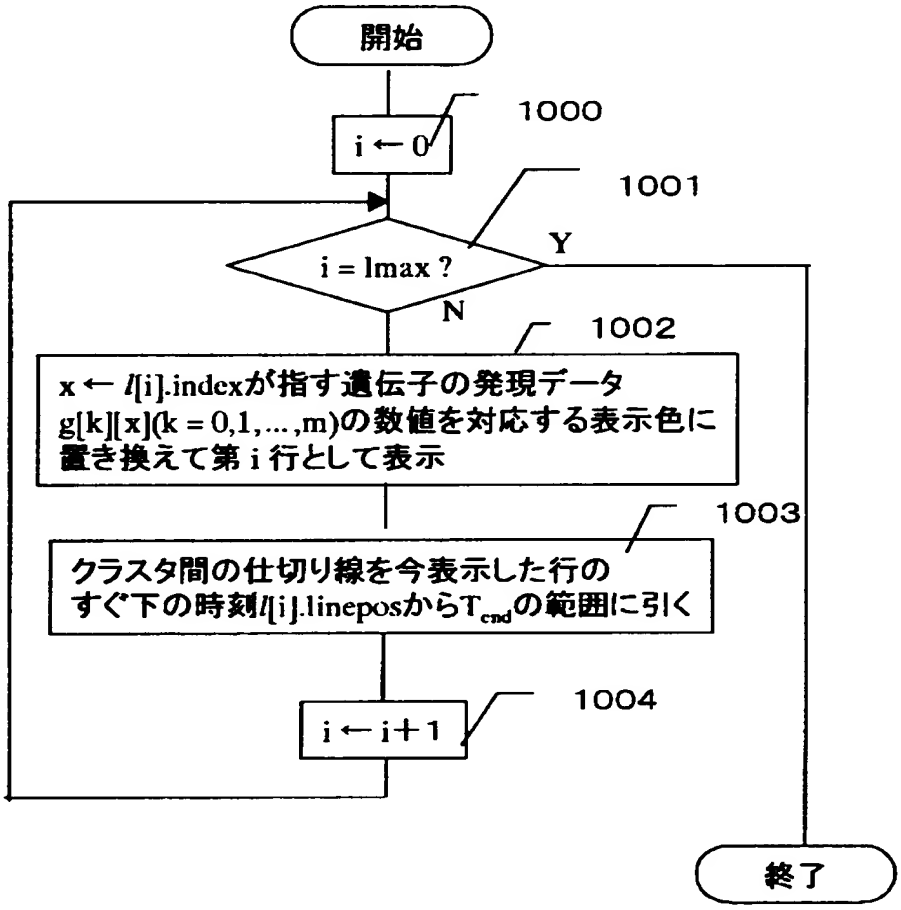
【図９】

図９


2
813
クラスタC[1]からC[ccnt]で、空集合でないものの総数をcmaxとする。
814
クラスタA[1]からA[cmax]を生成する。
815
空集合でないクラスタに対し、それに含まれる遺伝子の平均発現度の平均をとる。
すなわち、クラスタC[p] = {i1,..., ik}に対して、G'p= Gi1+...+ Gik / kを求める。
この値を降順に並べたものを、G'p1,..., G'pcmaxとしたときA[1] ← C[p1],...,
A[cmax] ← C[pcmax]を登録する。
816
総クラスタ数cmaxとクラスタA[1],...,A[cmax]を出力する
終了

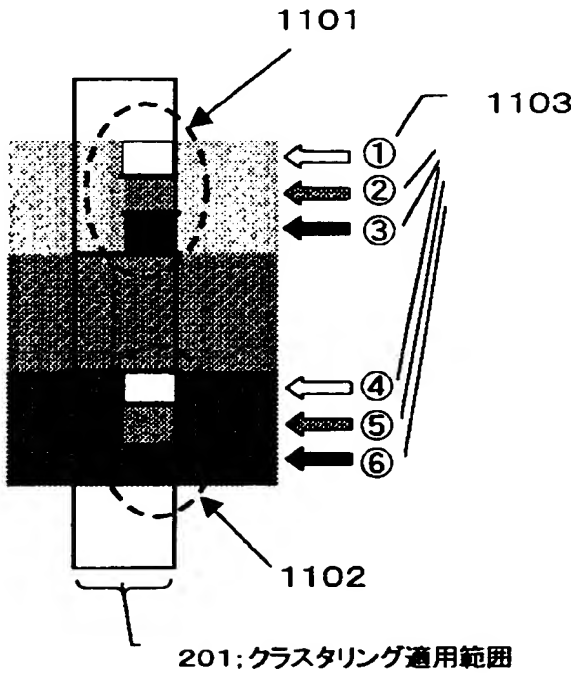
【図１０】

図10


開始
1000
i ← 0
1001
i = lmax ?
Y
N
1002
x ← l[i].indexが指す遺伝子の発現データ
g[k][x](k = 0,1,... ,m)の数値を対応する表示色に
置き換えて第 i 行として表示
1003
クラスタ間の仕切り線を今表示した行の
すぐ下の時刻l[i].linepos からT_end の範囲に引く
1004
i ← i+1
終了

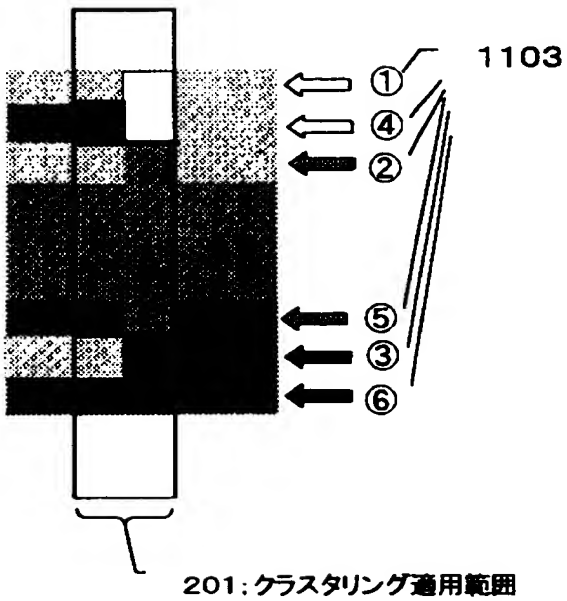
【図１１】

図11



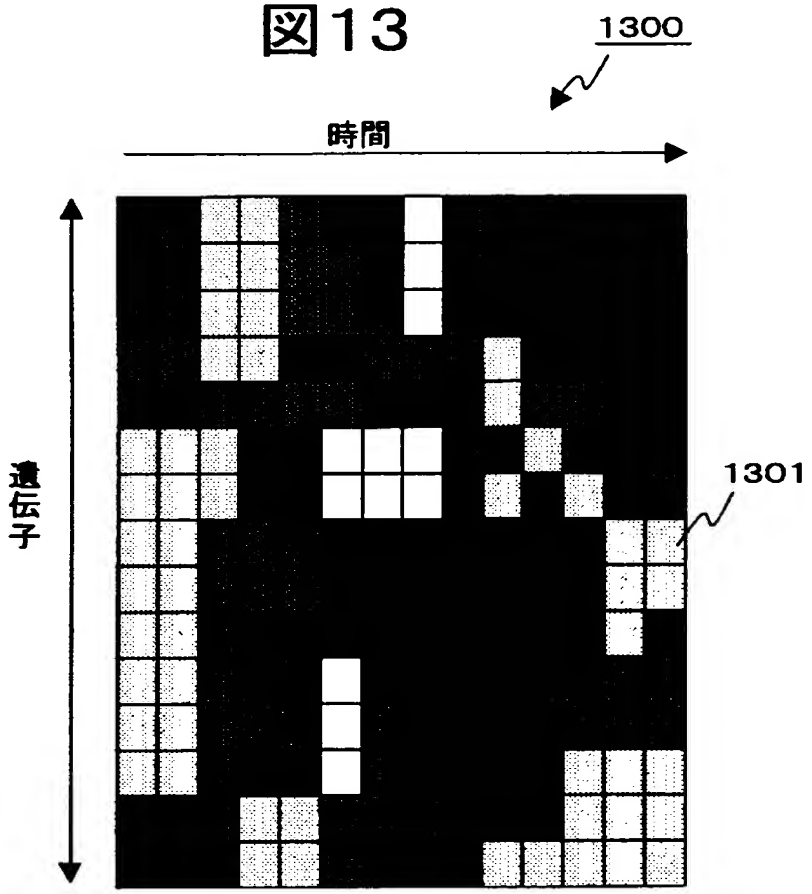
【図１２】

図12



【図１３】

図13



【書類名】　　要約書

【要約】

【課題】　ある時点において異なる遺伝子が同じ機能のために同様に発現しているが、ある時点では別々の役割を持つような場合を見つけ出し、効果的に表示すること。

【解決手段】　複数の遺伝子の時系列発現パターンデータの任意の時間区間を指定するステップと、指定された時間区間における時系列発現パターンデータを予め定めた基準値によってクラスタリングし、さらに同一クラスタ内において基準値を変えながら正または負の時間方向にクラスタリングを繰り返し行い、その結果を予め定めた表示形式で表示するステップとを備えることを特徴とする。

【選択図】　　図４

出 願 人 履 歴 情 報

識別番号　〔０００２３３０５５〕

１．変更年月日　１９９０年　８月　７日
［変更理由］　新規登録
住　所　神奈川県横浜市中区尾上町６丁目８１番地
氏　名　日立ソフトウエアエンジニアリング株式会社